图书情报工作
第66卷 第7期 2022年4月

基于文本挖掘的我国省级政府开放数据平台比较研究*

■ 陈美[1,2] 何祺[1,2]

[1] 中南财经政法大学公共管理学院 武汉 430073 [2] 中南财经政法大学国家治理与公共政策研究中心 武汉 430073

摘 要：[目的/意义]以我国14个省级政府开放数据平台为研究对象，从多个维度对其进行比较分析，为我国政府开放数据平台的发展提供参考建议。[方法/过程]通过爬虫技术获取数据，对数据进行描述性分析，并采用Tf-idf模型进行文本挖掘。以数据层维度和平台层维度为出发点，使用定性和定量分析方式，对数据资源细粒度、领域分布、时效性、格式种类、检索种类、访问转换率、用户反馈方面进行比较。[结果/结论]目前各省级开放数据平台发展程度不同，存在一定的改进空间，如应当结合本省(区、市)特点、数据集数量等综合考量数据集的发布方案，建设过程中需要注意开放平台数据检索方式、培训工作以及用户反馈等方面。

关键词：开放数据 政府开放数据 开放政府 比较

分类号：G250

DOI：10.13266/j.issn.0252-3116.2022.07.009

1 引言

政府开放数据平台是提高政府透明度的有效手段之一。通过对政府数据进行开放，“晒出”政府数据资产，既能保障用户公开获取政府数据的权利，从而提升对政府的信任，又能增加用户的参与、互动以及自我赋权。随着一些对个人或者企业有经济利用价值的数据得到开放，亦能推动社会经济的增长。尽管我国还未构建国家级的统一政府开放数据平台，但在2021年2月9日，国家信息中心发布《国家公共数据开放平台建设用户问卷调查通知》，广泛征求社会各界对各级开放平台的意见建议及对国家公共数据开放的具体需求，旨在进一步提升开放平台用户体验，促进数据供需衔接，释放更多数据开放红利[1]。这个通知的发布，对加快推动我国政府开放数据平台建设的进程具有重大意义，而地方政府开放数据平台是国家政府开放数据平台的前身，可以为国家政府开放数据平台建设提供经验参考，因而有必要加强对我国省市政府开放数据平台的研究。

从现有研究来看，主要采用定性和定量相结合的方法，从不同维度对不同国家、地区的开放数据平台进行对比，具体包括：①国家级政府开放数据平台。杨瑞仙等选取美国、英国、日本、澳大利亚等典型国家，分别从政策体系、保障机制和公开系统这3个方面进行比较[2]；吴钢和曾丽莹选取美国、英国、澳大利亚、加拿大以及我国建设较早的北京市和上海市政府开放数据平台作为调研对象，从资源现状、组织和检索、服务方式等方面探究当前国内外平台的发展现状[3]。②省市政府开放数据平台。谭必勇和陈艳以10个代表性东、中、西部省、市的开放政府数据平台为研究对象，对我国开放数据平台的质量进行研究[4]；余奕昊和李卫东以数据功能、接口功能、应用功能和互动功能4个方面为视角，对比10个省市政府开放数据平台的数据集数量、接口调用次数等数据，分析我国地方政府开放数据平台的现状和问题，并提出优化对策[5]。③城市政府开放数据平台。邓胜利和夏苏迪从数据层和平台层出发，以中美8个城市为例，将各地区政府开放数据平台资源数量、访问量、发布时间占比等进行对比，指出打造数据开放平台应当重点关注民生数据的开放、用户体验的优化等内容[6]。从研究方法来看，现有研究多采用比较分析、文献分析等定性研究和描述性分析相结合的方法，而较少侧重文本挖掘的方法。笔者在不同维度的对比基础上，采用Tf-idf文本挖掘模型进行定量分析，力图通过文本内容来反映出更为真实、准确

* 本文系国家自然科学基金项目“面向用户的开放政府数据使用行为机理及隐私风险控制研究”(项目编号：72004056)和中南财经政法大学中央高校基本科研业务费专项资金资助项目“开放政府数据政策优化研究”(项目编号：2722022BQ039)研究成果之一。

作者简介：陈美，副教授，博士，E-mail：chenmei672236@163.com；何祺，硕士研究生。

收稿日期：2021-11-11 修回日期：2022-01-17 本文起止页码：88-98 本文责任编辑：易飞

陈美，何祺. 基于文本挖掘的我国省级政府开放数据平台比较研究[J]. 图书情报工作,2022,66(7):88－98.

的情况。从研究对象来看,已有研究大多关注国家级和城市政府开放数据平台,虽然也有研究关注省级政府开放数据平台,但这些研究是将省级和市级政府开放数据平台放在一起进行比较,而较少专门针对省级政府开放数据平台进行比较。由于省级和国家级、市级的开放数据平台所对应的行政层级不一样,而且这些平台的数据资源的数量、平台规模也不同,前人研究的结论在各省级开放数据平台的适用程度有待商榷。因此,有必要对各省级政府开放数据平台进行专门比较,从中总结成功的经验,发现可能存在的问题并加以优化,为后续我国地方政府开放数据平台的发展和国家政府开放数据平台构建提供参考。

2　比较分析框架

笔者借鉴邓胜利和夏苏迪中的分类视角,从数据层、平台层两个维度展开对比[6]。数据资源细粒度、领域分布、时效性、格式种类能在一定程度上反映政府开放数据平台开放的广度和深度,作为描述数据层的指标;数据的检索种类、访问转换率和用户反馈则能反映用户与平台的互动交流的情况,作为描述平台层的指标。政府开放数据平台具体对比框架见表1。需要说明的是,二级指标中的数据资源包括数据集、APP、API。其中,数据集是指原始数据经过加工之后得到的包含数据的集合,其数量直接反映开放数据平台的发展程度。APP是开放数据平台中应用程序模块,APP数量越多,平台数据集的可用性越强。API是开发人员对政府数据进行调用的接口,API数量越多,平台数据集的开放程度越高,数据资源的价值发挥得越充分。因此,笔者采用数据集、APP、API三者综合反映数据层中数据资源。

表1　政府开放数据平台比较分析框架

一级指标	二级指标	三级指标	说明
数据层	数据资源细粒度	数据集细粒度	
		APP 细粒度	
		API 细粒度	
	领域分布	—	工业农业/教育文化/安全生产等
	时效性	—	
	格式种类	—	
平台层	检索种类	—	高级检索/关键词检索等
	访问转换率	—	
	用户反馈	—	文本挖掘模型计算出的高频词

3　样本选取与研究方法

笔者以复旦大学数字与移动治理实验室于2021年10月发布的《中国地方政府数据开放报告(指标体系与省域标杆)》中的18个省级评估对象为参考,并于2021年9月10－15日对平台数据进行采集。通过逐一访问这18个省级政府数据开放平台进行筛选,根据网站的有效性和数据采集的可行性,最终选取以下14个省(区、市)为研究对象,分别为湖南省、山东省、陕西省、江西省、宁夏回族自治区(以下简称“宁夏”)、河南省、浙江省、海南省、福建省、广东省、广西壮族自治区(以下简称“广西”)、贵州省、河北省、四川省(排名不分先后)。

需要注意的是,统计数据时可能存在以下情况:①平台的数据处于实时更新状态,各省(区、市)数据可能存在小范围的滞后性以及并非同天统计的情况;②统计时网页失效以及网页呈现方式为图表形式,导致本文各分析模块的研究省(区、市)数量可能存在稍许不同;③由于个别省(区、市)在某个分析模块数据记录数很少,为了方便后续梳理结论,在不影响结果的情况下部分省(区、市)将被去除。

笔者采用定性和定量相结合的研究方法,通过Python代码进行爬虫获取各省(区、市)的原始数据集记录、APP记录、API记录等原始数据,对数据层的各个指标以及平台层的检索种类指标、访问转换率进行对比分析,并利用Tf-idf模型挖掘用户在互动交流时关注的重点内容(见图1)。如此,从这些数据分析中发现问题,提出合理建议,为政府开放数据平台的完善提供可行方案。

4　各省级政府开放数据平台对比分析

4.1　数据资源细粒度

数据开放成为进一步研究和创新知识的第一阶段,对数据的生产、传播、管理和使用方式产生了直接的影响。合理划分数据资源,有利于降低用户获取数据时的时间成本和人力成本,提高政府开放数据平台的可用性。本文通过“细粒度”指标衡量数据资源划分的疏密程度,其计算公式为:数据资源细粒度＝资源数量/领域分布情况,具体采用数据集细粒度、APP细粒度、API细粒度进行综合反映。

由于同类数据资源数值相差较大,因而笔者采用数据资源细粒度的中值作为对比标准:如果高于标准,则认为数据资源划分较疏;反之,则划分较密。根据以上分析原则统计了14个省(区、市)的数据集、API接口、APP细粒度情况并进行对比(见表2),并将政府开放数据平台划分的疏密情况进行可视化,见图2。

从图2可以看出,我国各省级政府开放数据平台细粒度并不均衡,差异较大,大多数划分不太合理。

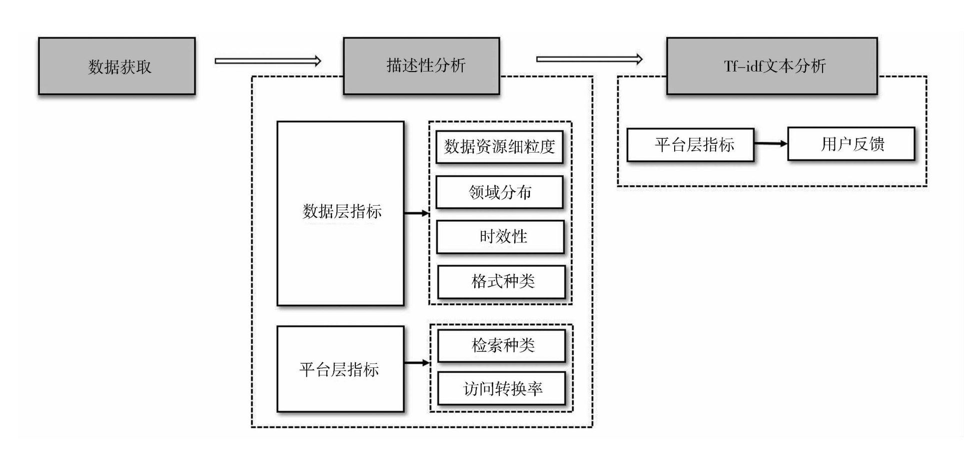


图 1 数据处理流程

表 2 数据集数量及涉及领域

省（区、市）	数据集			API 接口			APP		
	数量/个	领域数/个	细粒度	数量/个	领域数/个	细粒度	数量/个	领域数/个	细粒度
湖南	172	—	—	—	—	—	5	4	1.25
山东	940	19	49.47	66 379	19	3 493.63	150	19	7.89
陕西	141	10	14.1	—	—	—	9	5	1.8
江西	122	12	10.17	4	2	2	—	—	—
宁夏	1 770	38	46.58	8	4	2	3	2	1.5
河南	806	21	38.38	1 609	20	80.45	10	6	1.67
浙江	1 059	22	48.14	1 106	22	50.27	52	22	2.36
海南	233	16	14.56	1 256	22	57.09	—	—	—
福建	2 223	23	96.65	1 258	23	54.7	15	8	1.88
广东	6 182	12	515.17	235	11	21.36	54	11	4.91
广西	5 254	23	228.43	401	17	23.59	46	10	4.6
贵州	3 158	21	150.38	—	—	—	—	—	—
河北	209	10	20.9	—	—	—	—	—	—
四川	6 356	22	288.91	—	—	—	17	—	—

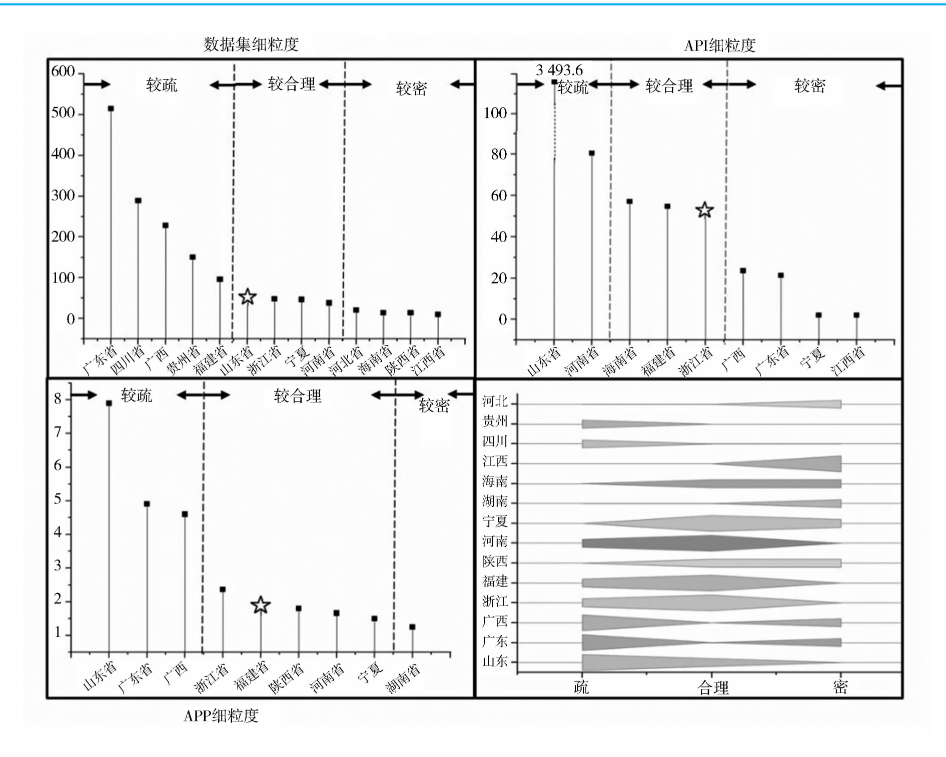


图 2 数据资源划分疏密程度对比

注：☆ 表示该省（区、市）数值居于中位数

陈美，何祺. 基于文本挖掘的我国省级政府开放数据平台比较研究[J]. 图书情报工作,2022,66(7):88－98.

其中,划分较疏的省(区、市)需要重新考虑分类,将较大概念的领域进行二次分类,重新界定归属。但是,数据集数量过大,会导致重新分类时实施困难,因而可考虑仅对新增的数据集进行加工。我国政府开放数据平台处于建设完善中,数据资源需求可能还不稳定,后续增幅未知。因此,划分较密的省(区、市)可维持现状,随着各省(区、市)数据集进入稳定发展阶段后,再考虑是否需要重新分类。

4.2 领域分布

数据集的领域分布指归属不同领域的数据集占总数据集的比例。结合各省(区、市)的经济水平、教育水平、政策风向等方面,不同领域数据集的占比可以反映出不同省级政府开放数据平台在该时期的工作要点和关注重心,不仅能深化用户对政府开放数据平台的理解,也可以为后续政府开放数据平台的发展导向提供有效的反馈。

经过汇总发现,14 个省级政府开放数据平台涉及的领域共计 79 个。由于其划分数据集的标准不一,可能会出现以下问题:①存在相近的领域,如"文化休闲"与"文化";②表达内容相同但表达的术语不同,如"安监"和"安全监管";③"暂无""其他"领域的含义不明确。针对上述问题,笔者采用以下解决方式:①将内容相似或者相同的领域进行合并,减少冗余的领域数;②将"暂无""其他"统一归为"未分配"领域一类。经过整理,将 79 个领域重新归类并划分为政务领域以及资源、能源、环境领域等 30 个领域,领域的分类和占比如表 3 和图 3 所示:

表 3 领域分布归类

一级指标	二级指标
政务领域	综合政务
资源、能源、环境领域	资源能源、资源环境、环境、资源、能源安全
质量领域	质量
知识产权领域	知识产权
医卫健康领域	医疗卫生、卫生健康、卫生、医疗
信用领域	信用体系、信用服务、信用
文体休闲领域	文化休闲、文化体育、文化
统计服务领域	统计服务、统计
市场监管领域	市场监管、企业登记监管、市场监督
食药安全领域	食药安全、食品药品安全
生态环境保护领域	生态环境、生态环保、生态
商贸领域	商业服务、经贸工商、商贸流通
气象领域	气象服务、气象
工业、农业领域	农业农村、工业农业、农业
信息科技领域	科技创新、信息技术、科技
经济、金融领域	经济建设、财税金融、金融
教育领域	教育文化、教育科技、教育
交通领域	交通运输、道路交通、交通出行、交通
机构团体领域	机构团体
社会民生领域	生活服务、健康保障、社区治理、社会资源、社会民生、民生服务、社会救助、社会发展、社会保障、社保就业、就业、社保
海洋领域	海洋
海关口岸领域	海关口岸
公共服务领域	公共服务
旅游服务领域	旅游服务
法律服务领域	法律服务
闽台合作领域	闽台合作
地理领域	地理空间、地理
城乡建设领域	城乡建设、城市建设、城建住房
安全领域	安全生产、公共安全、安全监管、安监
未分配领域	暂无、其他

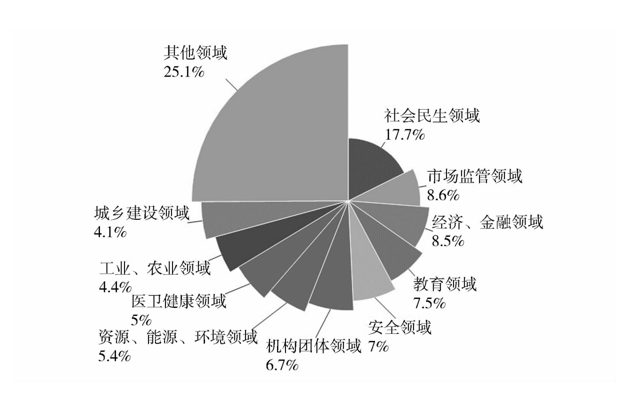


图 3　各领域数据集占比

已有文献表明,开放数据需要以用户的视角为切入点,以用户需求为导向进行体系框架的搭建[7-9],但数据本身反映的政治导向也是不能忽视的重要部分。近年来,引起广泛关注的热词,如"碳中和""双减""数字人民币""ESG""新冠肺炎"等持续刷屏,在一定程度上反映了我国政府目前的工作重心和方向。通过图 3 可知,政府开放的数据主要集中在民生、市场监管、经济金融、教育、安全、机构团体、资源、能源、环境、医疗卫生、工业农业、城乡建设等领域。由此可见,开放数据的内容与我国当前的战略发展方向实现了高度契合,不仅体现了开放数据的公共价值,也有利于在数据的可用性与用户需求之间实现良好平衡。

4.3　时效性

在数字驱动的全球背景下,时效性作为开放数据的原则之一,很大程度上决定着数据的质量。更重要的是,在以用户导向为理念的市场背景下,开放数据的时效性还能显著影响用户的满意度和对政府的信任。特别是在应对紧急情况时,时效性显得尤为重要[10]。相反地,"过时"的数据很大程度上代表一类"无效"数据,不仅实际意义不大,还可能会给数据库带来过载的困扰。因此,政府在进行管理时,一方面,要重视数据的时效性,及时更新数据清单目录;另一方面,也要注意对"过时"数据的备份和清理,减少系统空间的存储压力,从技术层面提高平台运行效率。各省级政府数据平台的时效性如表 4 所示:

表 4　各省级政府数据平台时效性

省(区、市)	区间	数据集总量/个	2017 年占比/%	2018 年占比/%	2019 年占比/%	2020 年占比/%	2021 年占比/%	趋势图	趋势说明
四川	[5 000, +∞]	6 356	0.00	0.00	6.36	39.84	53.79		持续上升
广东	[5 000, +∞]	6 182	2.98	31.90	21.48	15.88	24.13		
广西	[5 000, +∞]	5 254	0.00	0.00	0.00	45.98	54.02		
贵州	[500,5 000]	3 158	0.06	20.27	16.12	36.16	27.39		先上升后下降(不含特例)
福建	[500,5 000]	2 223	0.00	0.00	0.00	73.23	26.77		
宁夏	[500,5 000]	1 770	0.00	0.00	0.00	0.06	99.94		
浙江	[500,5 000]	1 059	0.00	0.00	42.97	33.90	23.04		
山东	[500,5 000]	940	21.17	14.57	1.28	46.49	16.49		
河南	[500,5 000]	806	0.00	52.73	47.02	0.25	0.00		
海南	[0,500]	233	0.00	50.64	6.44	11.16	31.76		整体保持上升(不含特例)
陕西	[0,500]	141	4.26	83.69	4.26	4.96	2.84		
江西	[0,500]	122	16.39	13.11	0.00	0.00	70.49		

由表4可知,各省级政府开放数据平台的时效性差异较大,发展水平各不相同。具体而言,在数据集数量在[5 000,+∞]的省(区、市),每年发布的数据集数量整体呈现持续向上增长的态势;在数据集数量在[500,5 000]的省(区、市),每年发布的数据集数量呈现先上升后下降的趋势;在数据集数量在[0,500]的省(区、市),每年发布的数据集数量整体保持上升趋势。之所以前期上升,可能是因为平台处于初始建立时期,数据集的发布经历了从"0－>1"的过程,因而趋势上表现为持续上升;出现下降可能是因为平台进入稳定运行阶段,导致增幅会有所下降等。需要注意的是,由于宁夏政府开放数据平台建立较晚,虽然数据集总量居中,但仍呈现上升趋势;陕西省经过2018年的快速发展后,可能因为陕西省关注重点偏向公共服务和信用领域,而归属此类领域的数据集大部分属于年度数据,导致后续数据集发布数量很少。因此,在平台稳定运行后,即使数据集总量不大,数据集数量也会呈现上升之后又下降的趋势,而中间的拐点一般在每年年末。

4.4 格式种类

政府开放数据平台的数据集正持续不断地增长,而数据集往往以相应格式进行存储。格式种类越丰富,政府数据平台的开放程度越高。从图4和图5中可以看出,平台数据格式以XLSX、JSON、XML、CSV为主,格式种类在5－8种左右为宜。结合数据集总量来看,整体分布较为合理。具体来看,海南省数据集仅有233个,因而只选用XLS这1种使用最普遍的数据集格式。广东省和四川省的数据集数量相差不大,但数据集格式的种类数却成倍数关系。这是因为,广东省数据集格式中PDF、TXT、DOC 3种数据集格式只有个位数,数量过少,因而这3种格式可忽略不计。另外,在对格式种类进行规划时,尽可能避免类似格式都被全部采用,如"DOCX"和"DOC""XLSX"和"XLS"等,避免给系统增加不必要的运行负担。

4.5 检索种类

按照世界银行的界定,只有同时满足以下两个条件的数据才被认为是开放:①合法开放,以允许商业与非商业使用和无限制重复使用的方式明确许可;②技术上开放,以机器可读的标准格式提供,这意味着它可以被其他常用的计算机应用程序检索和有意义地处理。因此,数据检索种类是体现政府开放数据的开放程度高低的途径之一[1]。数据检索种类配置不合理,对于用户来说,会影响搜索效率,降低使用满意度,甚

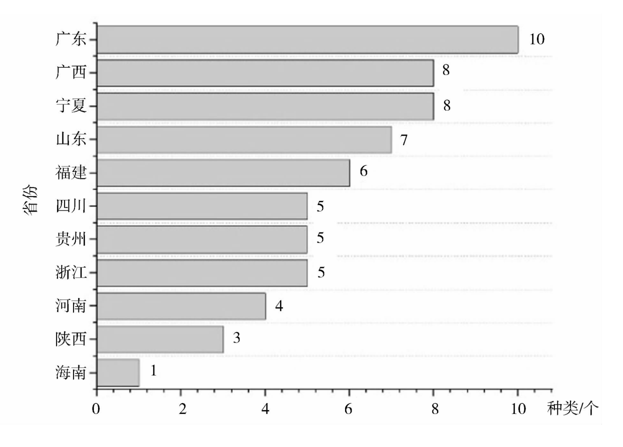


图4　数据集格式种类数

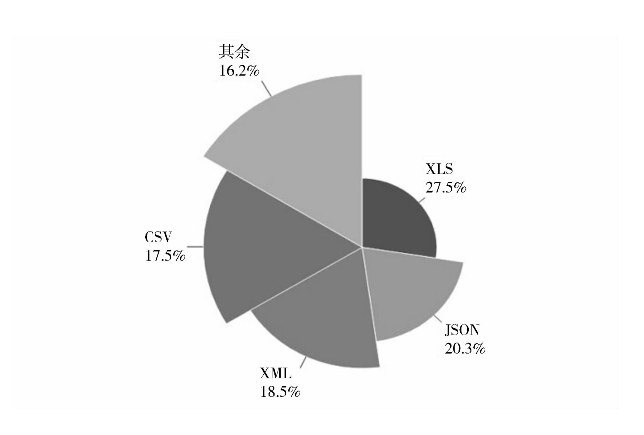


图5　数据集格式占比

至无法查询获取需要的数据;对政府来说,可能会增加平台的开发维护成本和人力成本、降低系统的性能。为进一步进行分析,笔者将各省级政府数据平台的检索情况进行汇总,见表5。

由表5可知,关键词检索、目录检索和其他检索中的管辖部门检索、领域检索、地图服务、数据格式检索、地点检索、时间范围检索、开放方式检索为常见检索方式。数据集数量大于800的省(区、市),其数据检索种类基本都在8种以上;数据集数量小于800的省(区、市)数据集检索种类基本在4－5种。因此,数据检索方式的选择应当结合数据集数量来确定。另外,部分省(区、市)还采取多种检索方式并行的方式,利用不同检索方式的优势进行互补、搭配使用,以实现检索效率的最大化,这也值得参考借鉴。

4.6 访问转换率

访问和下载是一种用户行为,是用户与平台进行交互的一种方式。访问可能仅仅说明用户对该领域的关注,而下载更能真实反映用户对数据集的需求。因此,结合访问量和下载量,用其比值访问转换率则更能全面衡量数据集的"热度",反映数据集的"吸引力",即访问转换率＝下载量/访问量。

表 5 各省级政府数据平台的检索情况汇总

检索方式		省(区、市)/数据集总量/个														
		四川 6 356	广东 6 182	广西 5 254	贵州 3 158	福建 2 223	宁夏 1 770	浙江 1 059	山东 940	河南 806	海南 233	河北 209	湖南 172	陕西 141	江西 122	合计 28 625
文本检索	关键词检索	√	√	√	√	√	√	√	√	√	√	√	×	√	√	13
	高级检索	√	×	√	×	×	√	√	√	×	×	×	×	×	×	5
目录浏览检索		√	√	×	×	√	√	√	√	√	√	√	√	√	√	12
其他检索	管辖部门检索	√	√	×	√	√	√	√	√	√	√	√	√	√	√	13
	领域检索	√	√	√	√	√	√	√	√	√	√	√	×	√	√	13
	地图服务	√	×	√	×	√	√	√	√	√	×	×	×	√	√	9
	数据格式检索	√	√	√	√	×	√	√	√	√	×	×	×	√	×	9
	地点检索	√	×	√	×	√	√	×	√	√	×	×	√	×	×	7
	时间范围检索	√	×	√	×	×	√	×	√	×	×	×	×	√	×	5
	开放方式检索	√	√	√	×	×	√	×	√	×	×	×	×	×	×	5
	接入方式检索	×	×	√	×	×	√	×	×	√	×	×	√	×	×	4
	搜索词位置检索	√	×	×	×	×	×	×	√	×	×	×	×	×	×	2
	开放机构检索	×	×	×	×	×	×	×	×	×	×	×	×	√	×	1
	文件类型检索	×	×	×	√	×	×	×	×	×	×	×	×	×	×	1
	摘要检索	×	×	×	×	×	×	√	×	×	×	×	×	×	×	1
	评分检索	×	×	×	×	×	×	×	×	×	√	×	×	×	×	1
合计		11	6	9	5	6	11	8	11	8	5	4	4	8	5	

访问转换率不仅仅是“线性”反映数据集价值高低的指标,更体现着一种“螺旋式”的隐性反馈机制。政府根据数据集的访问转换率来判断用户需求,有助于其把握开放数据平台发布内容的重心和方向,换句话说,即政府可以通过用户对数据集的访问和下载情况,来验证政府对用户需求的判断是否准确。若存在偏差,则政府需要进行及时调整。访问转换率越高,则说明数据集越符合用户需求。在实际应用中,可能会产生访问转换率“虚”高的情况,如下载量和访问量均处于较低水平,此时可通过去除极值或事后校对审核的方式进行数据处理,筛选出有意义的数据进行分析。由此看来,对于政府开放数据平台评估而言,访问转换率不仅是一个比较重要的参数,也是政府在开放数据的探索过程中用来及时调整战略方向的风向标。

笔者分别用访问量排名和下载量的排名对访问量和下载量进行衡量。从图 6 可以看出,各领域内数据集数量排名和访问量排名、下载量排名成线性关系,访问量排名和下载量排名保持一致。换言之,数据集数量越多的领域,访问量和下载量越多,而且用户访问数据集时通常都会进行下载。这些数据质量较高,基本达到了内容符合用户需求的标准。但是,部分省(区、市)仍需要进行适当改进,如福建省访问量和下载量在一致性方面欠佳,说明虽然用户对该领域数据感兴趣,但数据内容质量方面达不到用户需求,因而用户没有进行进一步的下载操作。又如,海南省数量较少的数据集反而拥有较高的访问量和下载量,说明政府还未充分站在用户需求的角度去考虑数据集的发布,没有及时做好数据集的追踪,造成了数据集发布方向不明确,使得用户希望获取的数据反而被忽略。

4.7 用户反馈

政府开放数据平台的建立大大增加了用户参与公共事务的机会。一般来说,普通用户与公共事务建设之间的“不良连接”,忽略了用户体验潜在的价值。但是,通过在政府开放数据平台创建互动交流栏目,能够克服这个缺点。具体而言,可以通过开放数据平台上互动的强度、互动方式的丰富性等来改善连通性,获得用户反馈。笔者将用户反馈视为开放数据使用后的体验情况以及用户的数据需求,包括使用过程中出现的数据问题、系统功能问题、提出的改善建议、数据需求申请等。

笔者将政府开放数据平台的子栏目下的互动文本数据进行归类,分为“数据申请”“问题纠错”“意见反馈”“咨询问题”四大模块。具体划分见表 6。

基于上述划分的模块 - 子栏目划分,笔者使用 Tf-idf 算法对各栏目下的用户反馈内容进行文本挖掘分析。通过哈工大停用词表,附加自定义词语(主要是标点符号、官网回复用户时的礼貌用语以及省(区、市)名称),利用 Jieba 进行分词,得到排名前 20 的词语后,

陈美，何祺.基于文本挖掘的我国省级政府开放数据平台比较研究[J].图书情报工作,2022,66(7):88－98.

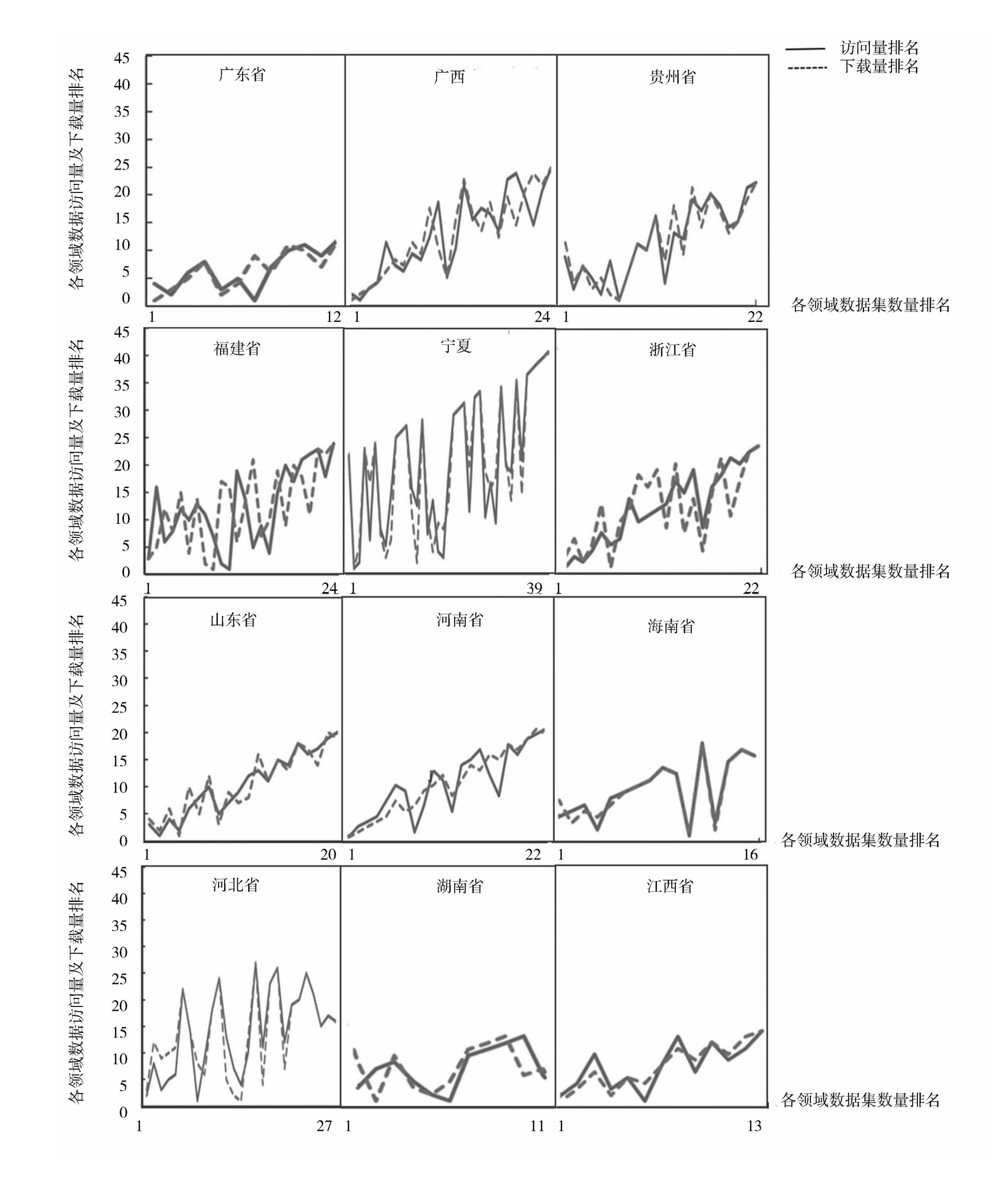


图6　各领域数据访问量、下载量排名

表6　模块－子栏目

模块	子栏目	作用
数据申请	数据申请、需求列表、需求申请	反映出用户的数据需求
问题纠错	纠错公开、数据纠错、纠错列表、数据问题	完善政府的数据治理
意见反馈	意见征集、平台建议、内容建议、平台体验	体现用户的关注焦点
咨询问题	常见问题、问题反馈、咨询提问、问题咨询	改进平台功能

按不同类目(标题词频、问题描述、回复词频)绘制词云图进行可视化,得到主题词并追溯相应原文,以便更准确地对主题词加以解读。

Tf-idf算法的目的是评估词语的重要性。其基本原理主要是词语的重要性与词语在文本中出现的次数成正比,并与其在语料库出现的次数成反比。换言之,某词在文本中出现频率很高,在其他文章中出现的频率也很高,那么它的重要性并不很大;但若在文本中出现频率很高,而在其他文章中出现的频率很低,则说明重要性很大。该算法优势在于:可以较好地过滤无意义或者不相关的词语(如“通知”“青岛”“那么”等),提高文本筛选后内容的真实性,而且简单、快速。虽然它无法区分一词多义的情况,但是鉴于文本来源于相对客观的陈述文本,并非带有感情色彩的叙述类文本,最终本文选择该方法进行高频词提取。最终得到的各模块关键词词云图,见图7－图9。

对于用户而言,反馈的建议得到采纳并解决,能极大地提升其参与积极性。这会有利于形成良性的反馈循环,整体上为我国政府数据开放平台后续的发展奠定坚实的基础。



图 7 “标题词频”模块词云图

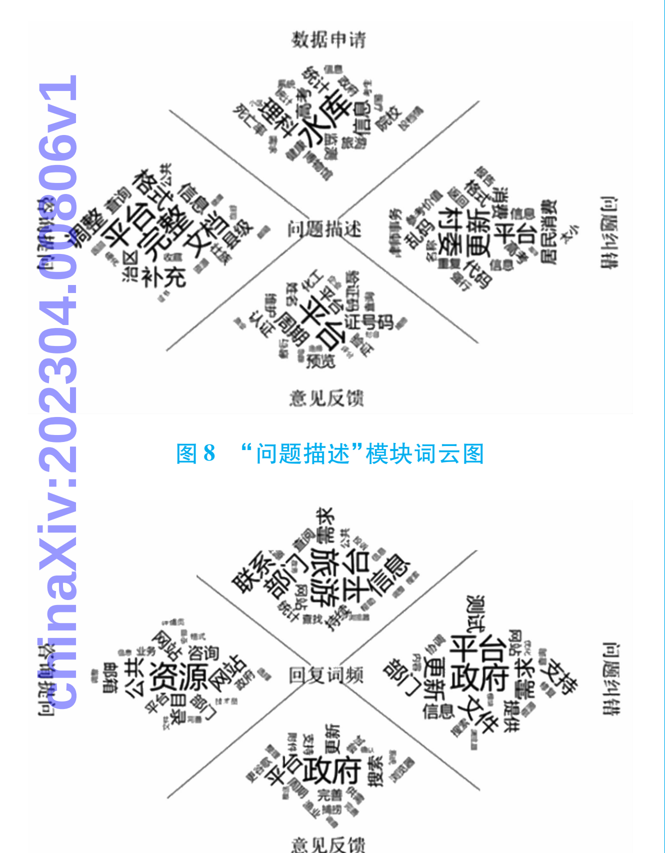


图 8 “问题描述”模块词云图

图 9 “回复词频”模块词云图

5 总结与建议

5.1 数据层

5.1.1 数据领域的划分

我国政府开放数据平台数据集领域的划分合理性较差，在完善时应当结合数据集数量、现行政策以及本省（区、市）特色综合考虑，不应划分得过疏或过密，避免导致政府维护成本增加以及用户搜索困难等问题。

5.1.2 数据格式的选择

目前，各省级政府开放数据平台数据集格式配置较为合理。在后续平台建设时，数据集格式可以将 XLSX、JSON、XML、CSV 这 4 种作为基础格式，并且尽量避免使用类似功能的格式。其余格式则可根据需求、收集数据难易程度或数据本身情况等综合考虑是否进行发布。

5.1.3 数据内容的发布

经过对比分析，当前开放数据平台发布的内容与政策方向契合度较高，但仍存在一定的改进空间。一方面，数据集的发布应当结合访问转换率的指标进行考虑，对热点领域的数据集应当重点关注，访问量很少的数据集可以考虑不进行发布。由于用户的关注点是动态变化的，因而政府应当定期梳理追踪，做好数据集发布的清单目录，重视数据集发布的“靶向”性，提高平台数据的实用价值。另一方面，数据集的发布应当保证数据的“新鲜度”，及时更新，为用户提供最新的数据，并做好旧数据集的备份处理。这也是加强政府数据开放平台数据质量治理、提高用户对政府的满意度的途径之一。

5.2 平台层

5.2.1 检索方式的使用

信息检索应当具备全面性、灵活性、高效性特点。因此，各省（区、市）在加强对政府开放数据平台数据集管理的同时，应当注意结合该省（区、市）数据集的数量配置，对多种检索方式相互搭配使用。这样既能提高用户的检索效率，也能减少开发人员的维护成本。

5.2.2 培训工作的开展

由政府或非营利组织做好政府开放数据平台使用的培训工作，如通过抖音、微博等线上平台来宣传使用方法；编纂用户使用指南来引导用户操作等。例如，与北美和欧洲不同，非洲国家获得开放政府数据的途径往往有限，因而它们驾驭技术并将其作为增长动力的能力也有限。CFA（Code for Africa）是非洲最大的公民技术和数据新闻实验室网络，在 20 个国家或地区设有团队。这个组织旨在培养社区内的技术和编码技能，为公民创造机会，让他们成为政府、企业和公共机构的监督者。这个组织不仅将开放数据视为潜在的公共资产，而且已经开发了一个数据奖学金项目，能将数据技能培训的人员嵌入各种媒体和非营利组织的项目中。

5.2.3 用户反馈的完善

相关人员在对用户的反馈进行解答时,尽可能模板化、具体化,把切实解决用户的问题作为目标。空话、套话会让用户产生较差的体验感,从而对用户与政府之间的互动交流产生负面影响,不利于政府数据开放平台的长期发展,因此应尽量避免。基于词云图的主题词,笔者从中总结出未来平台应当改进的方向,大致有如下几点:

(1)内容精准、及时化。目前,用户在需求上更倾向获取关于教育文化、行业资源方面的数据信息,如高考分数、水库资源、旅游数据等。但是,用户的需求是动态的,意味着政府应当综合用户反馈、时事热点、国家大政方针等做出调整,并且及时更新。如此,才能让开放数据平台的数据真正得到利用,做到用数据驱动经济社会的发展。

(2)功能简洁、便捷化。政府开放数据平台部分功能使用存在技术问题,如实名认证操作异常等。但政府开放数据平台相关负责人员则回复,用户反映的这些问题中有部分是正常运行的。两者产生冲突的原因有可能是因为政府开放数据平台不稳定、浏览器不兼容、用户不会操作等。针对上述现象,笔者提出如下建议:首先,重视数据治理,提高数据的真实性、准确性;其次,通过多渠道宣传指导、对新增页面进行功能说明、优化界面、明确各栏目查询路径等方式,加强对用户的应用指导,减少用户操作过程中不必要的"弯路";最后,以用户为中心,考虑不同人群在使用上的难处,如新增老年模式、残疾人模式等专门页面,为弱势群体打开绿色通道,提升用户体验。

(3)解答专业、具体化。在解答用户反馈的问题时,"由于系统异常等原因,需要技术进行维护"等不太具体的回复不在少数。这种解答往往不仅不能解决问题,还可能起到反作用,让用户产生烦躁等消极情绪,并认为反馈是一种浪费时间的无用行为。因此,建议制定具体的解答模板,如问题描述、解决方式、解决时间、处理机构、举报邮箱等。更重要的是,政府应当做好反馈解答的检查工作,定期复查问题的解决情况。

6 结语

搭建政府数据开放平台的初衷是为了促进开放数据得到广泛的开发利用[15],增强用户对政府的信任,方便用户通过数据进行研究创新,推动社会的发展。笔者通过对各省级政府开放数据平台的数据层和平台层进行比较研究,发现能够借鉴的优势以及存在的问题,并提出了相关的建议。在研究过程中,由于政府开放数据平台的内容是动态更新的,使得本文所搜集的数据存在一定的滞后性,从而给分析结果带来影响。在今后的研究中,可以通过不同的研究方式或技术手段来解决此类问题。另外,政府开放数据平台比较的维度还有进一步扩展的空间,未来研究可以增加研究对象、扩展对比维度,从而提升政府开放数据平台发展建议的普适性。

参考文献:

[1] 国家公共数据开放平台建设用户问卷调查通知[EB/OL].[2021-10-25]. http://www.sic.gov.cn/News/612/10773.htm.
[2] 杨瑞仙,毛春蕾,左泽.国内外政府数据开放现状比较研究[J].情报杂志,2016,35(5):167-172.
[3] 吴钢,曾丽莹.国内外政府开放数据平台建设比较研究[J].情报资料工作,2016(6):75-79.
[4] 谭必勇,陈艳.我国开放政府数据平台数据质量研究——以十省、市为研究对象[J].情报杂志,2017(11):99-105.
[5] 余奕昊,李卫东.我国地方政府数据开放平台现状、问题及优化策略——基于10个地方政府数据开放平台的研究[J].电子政务,2018(10):99-114.
[6] 邓胜利,夏苏迪.中美城市政府开放数据平台对比研究[J].图书馆杂志,2019,38(6):57-68,75.
[7] 吴群英,马蕾.我国省级政府开放数据平台建设现状调查研究[J].情报探索,2020(9):69-75.
[8] 陈水湘.基于用户利用的政府数据开放平台价值评价研究——以19家地方政府数据开放平台为例[J].情报科学,2017(10):94-98,102.
[9] 汪庆怡,高洁.面向用户服务的美国政府开放数据研究及启示——以美国Data.gov网站为例[J].情报杂志,2016(7):145-150.
[10] 张林轩,储节旺,蔡翔,等.我国地市级政府数据开放发展现状及对策探析——以安徽省为例[J].情报工程,2021,7(4):79-92.

作者贡献说明:

陈美:研究设计、撰写论文;
何祺:搜集数据、撰写论文。

A Comparative Study on Open Data Platforms of Provincial Government in China Based on Text Mining

Chen Mei[1,2] He Qi[1,2]
[1] School of Public Administration, Zhongnan University of Economics and Law, Wuhan 430073
[2] National governance and Public Policy Research Center, Zhongnan University of Economics and Law, Wuhan 430073

Abstract: [**Purpose/Significance**] Taking 14 provincial government open data platforms in China as the research object, this paper makes a comparative analysis of them from multiple dimensions, providing references and suggestions for the development of government open data platforms in China. [**Method/Process**] The crawler technology was used to acquire data, and the descriptive analysis of the data was carried out, and the Tf-idf model was used for text mining. Starting from the dimensions of data layer and platform layer, qualitative and quantitative analysis methods were used to compare fine granularity of data, domain distribution, timeliness, type of format, type of retrieval, access conversion rate and user feedback. [**Result/Conclusion**] At present, open data platforms in different provinces(autonomous regions and municipalities) have different degrees of development, and there is certain room for improvement. For example, the release plan for data sets should take into account the province(autonomous regions and municipalities) characteristic and the number of data sets, etc. In the process of construction, attention should be paid to the open platform data retrieval methods, training and user feedback.

Keywords: open data government open data open government comparison

《图书情报工作》2022 年重点选题指南

1. 国家重大战略需求与图情档研究的作用与能力
2. 图书馆学、情报学、档案学研究方法与技术创新
3. 开放科学环境下科学交流范式的新变革
4. 后疫情时代学术信息交流模式的变化与影响
5. 新时代“信息资源管理”学科内涵与理论体系构建
6. 新文科建设视角下“信息资源管理”学科战略规划
7. 科技竞争背景下国家文献资源保障策略研究
8. 全媒体数字资源中心的设计与研究
9. 政府数字资源管理与长期保存
10. 政府开放数据管理与隐私保护
11. 开放科学数据、数据安全与个人信息保护
12. 数字经济中的数据功能及作用机制
13. 面向深度知识服务的拓展型信息资源标准与规范研究
14. 基于数据挖掘的文献资源智能采选推荐算法研究
15. 面向高价值专利培育的知识产权信息服务理论研究与实践探索
16. 面向交叉学科的跨学科知识组织方法与实践研究
17. 国内外情报工作制度演变与我国情报工作制度创新
18. 支持高水平科技自立自强的情报学理论方法
19. 关键核心技术重大突破情报监测与识别理论与方法
20. 聚焦创新驱动的核心关键领域情报分析服务研究
21. 面向国家发展战略需求的安全情报研究
22. 中美科技对抗下国家情报战略研究
23. 智能情报与数据智能研究
24. 国家总体安全观下应急管理信息服务及情报体系
25. 重大突发事件下应急情报协同及舆情引导
26. 高校图书馆在履行高校五大基本职能过程中的作用研究
27. 智慧图书馆研究与应用实践创新
28. 图书馆高质量发展的内涵与评价
29. 图书馆大安全管理与应急服务
30. 国际图书馆管理与服务发展趋势研究
31. 图书馆多源数据融合及治理
32. 图书馆小数据与暗数据的价值评估与应用研究
33. 教育新基建与图书馆建设
34. 图情档机构重组与队伍的专业化研究
35. 图书馆助力乡村振兴的策略研究
36. 健康信息学的理论与方法
37. 健康信息行为和个人健康信息管理
38. 虚假健康信息治理
39. 风险信息的识别、监测与传播
40. 区块链与信息安全问题
41. 面向全民全社会的数字素养能力与数字素养教育
42. 图书情报与档案管理学科课程思政建设
43. “元宇宙”场域下图情档学科的研究课题设置
44. 文旅融合背景下图书馆与档案馆服务创新
45. 数字人文与数字学术的新发展
46. 面向文化遗产的数字人文研究
47. 少数民族文献遗产建档研究
48. 红色文献、红色档案与红色记忆研究
49. 档案治理能力提升研究
50. 档案计算学
51. 数字出版与新型出版研究
52. 学术评价改革与创新
53. 数智赋能的创新评价
54. 其他

《图书情报工作》杂志社
2021 年 12 月